BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e137799
doi: 10.3897/biss.8.137799

OPEN ACCESS

Conference Abstract

# PyDwCA: A Tool for Integrating Biodiversity Data

Juan M. Sáez-Hidalgo[‡], Ricardo A. Segovia[‡,§], Francisco A. Squeo[‡,|], Pablo C. Guerrero[‡,§,¶]

‡ Institute of Ecology and Biodiversity, Concepción, Chile
§ Departamento de Botánica, Universidad de Concepción, Concepción, Chile
| Universidad de La Serena, La Serena, Chile
¶ Millennium Institute Biodiversity of Antarctic and Sub-Antarctic Ecosystems, Santiago, Chile

Corresponding author: Juan M. Sáez-Hidalgo (jmsaez@ieb-chile.cl), Ricardo A. Segovia (rsegovia@ieb-chile.cl)

Received: 25 Sep 2024 | Published: 25 Sep 2024

Citation: Sáez-Hidalgo JM, Segovia RA, Squeo FA, Guerrero PC (2024) PyDwCA: A Tool for Integrating Biodiversity Data. Biodiversity Information Science and Standards 8: e137799.
https://doi.org/10.3897/biss.8.137799

## Abstract

The Darwin Core Archive (DwC-A) format, based on the Darwin Core standard (Wieczorek et al. 2012), facilitates the exchange, management, and integration of biodiversity data from multiple sources. This ability to collate biodiversity data allows datasets to be aggregated at community-supported infrastructures, merged in different combinations, meta-analyzed and submitted to public repositories (Baker et al. 2014). Thus, the DwC-As serve as unifying archives in concatenated collective efforts, such as biodiversity inventories at different spatial and taxonomic scales.

Here we describe PyDwCA*[1], [2], a Python library implemented to handle the "star scheme" of DwC-A. This new library reads compressed zip files containing the expected meta.xml and uses it to assign the core component and its extensions. It also provides Python classes to define the core, the extensions, and the metadata file for creating an archive and writing it into a compressed zip file. PyDwCA also implements functionality to select, filter and merge DwC-A files.

We present this new tool in the context of the construction of the Chilean National Biodiversity Inventory (Fig. 1), but PyDwCA serves as a versatile technical solution applicable to different contexts in the field of biodiversity informatics (e.g., integration of datasets from biological collection and sampling events). To exemplify how PyDwCA works, we present the step-by-step integration of the Chilean Catalogue of Vascular Plants (Rodriguez et al. 2018) on a matrix provided by the Catalogue of Life (Banki 2024

), filtered with the species with occurrences recorded for Chile in the Global Biodiversity Information Facility (GBIF) (GBIF.Org 2023).

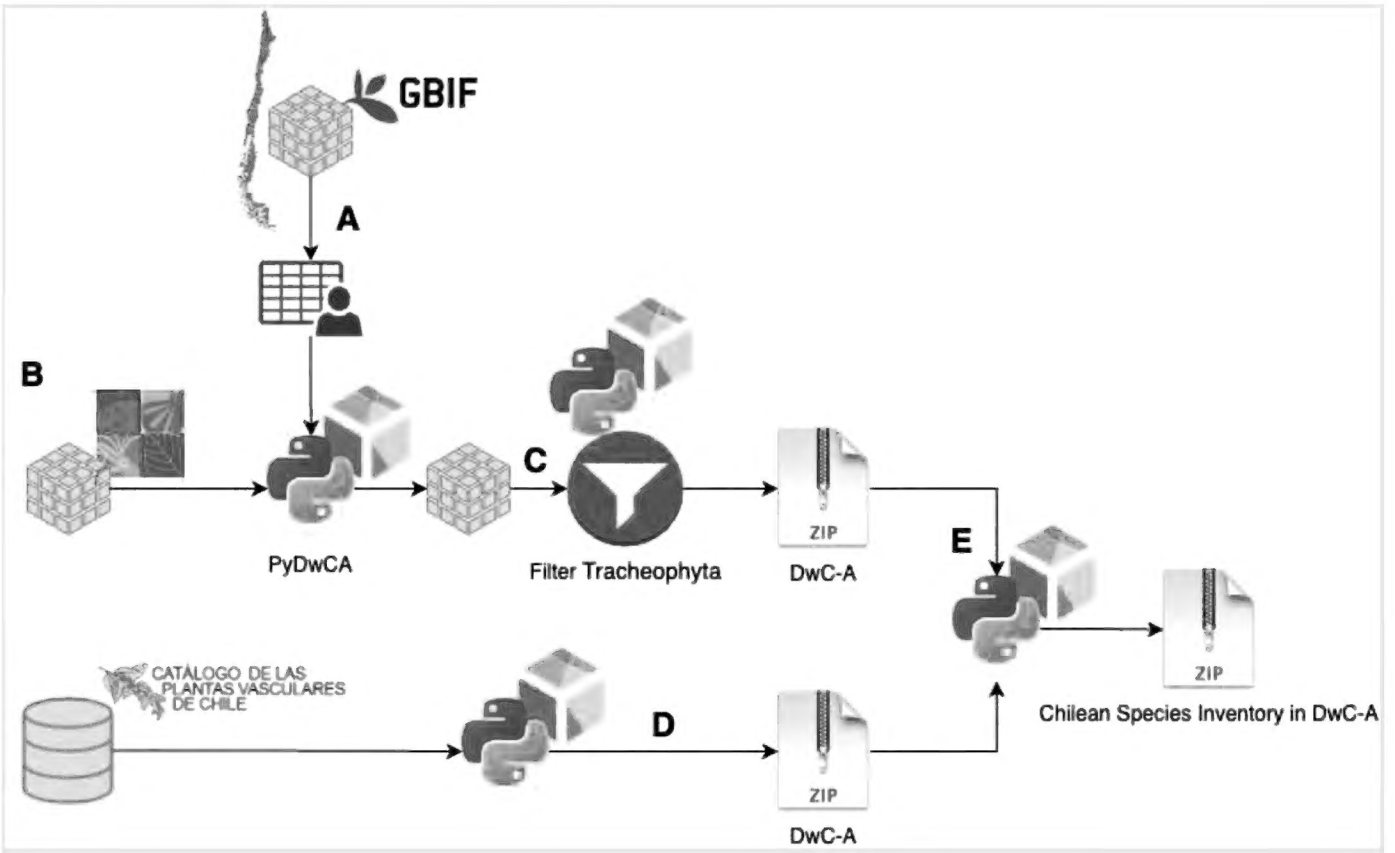


Figure 1.

Data pipeline for the generation of the Chilean National Biodiversity Inventory. A) Acquisition of the species presented in Chilean territory using the GBIF data platform. B) Download the DwC-A of the Catalogue of Life, filtering the species using the list obtained by GBIF and the PyDwCA library. C) Exclusion of species of Tracheophyta using the package. D) Generation of the DwC-A of the Catalogue of Vascular Plants of Chile using the Python library presented. This contains a curated list of the species of Tracheophyta in Chile. E) Merging of both DwC-A to get the first version of the Chilean National Biodiversity Inventory.

## Keywords

Darwin Core Archive, Python, taxonomic inventory

## Presenting author

Juan M. Sáez-Hidalgo

## Presented at

SPNHC-TDWG 2024

## Funding program

Centros Científicos y Tecnológicos de Excelencia con Financiamiento Basal; Institutos MILENIO

## Grant title

Centro Basal Instituto de Ecología y Biodiversidad (ANID FB210006); Milenio BASE (grant ICN2021_002)

## Hosting institution

Instituto de Ecología y Biodiversidad

## Conflicts of interest

The authors have declared that no competing interests exist.

## References

- Baker E, Rycroft S, Smith V (2014) Linking multiple biodiversity informatics platforms with Darwin Core Archives. *Biodiversity Data Journal* 2 (e1039). [In english]. https://doi.org/10.3897/BDJ.2.e1039
- Banki O, et al. (2024) Catalogue of Life Checklist (Version 2024-03-26). Catalogue of Life https://doi.org/10.48580/dfz8d
- GBIF.Org (2023) GBIF Occurrence Download. https://doi.org/10.15468/DL.QBUE7X. Accessed on: 2023-7-24.
- Rodriguez R, Marticorena C, Alarcón D, Baeza C, Cavieres L, Finot V, Fuentes N, Kiessling A, Mihoc M, Pauchard A, Ruiz E, Sanchez P, Marticorena A (2018) Catálogo de las plantas vasculares de Chile. Gayana. Botánica 75 (1): 1-430. https://doi.org/10.4067/s0717-66432018000100001
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715

## Endnotes

*1 PyDwCA library main page https://pypi.org/project/pydwca/

*2 PyDwCA GitHub repository https://github.com/IEB-BIODATA/pydwca